GRAMENER CASE STUDY SUBMISSION

Group Members:

1. Ranjidha Rajan - (DDA1710198)
2. Soumadiptya - (DDA1710089)
3. Nihar Behara - (DDA1710367)
4. Anugraha Sinha - (DDA1710381)

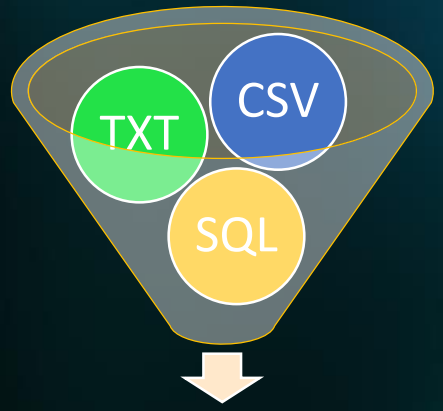
Loan Application – Evaluation Objective

- Business Objective
  - Primary Objective
    - Identify "High Risk" loan applicants
  - Secondary Objective
    - Use EDA to understand driving factors behind loan default
  - Minimize Risk
    - Not approving loans leading to a loss of Business
    - Approving risky loans leading to Financial losses
  - Make Informed Decision
    - Reject Loans
    - Reduce the amount of loan
    - Lend at higher interest rates

- Some Insights to Data
  - 3 kinds of loan categories covered
    - Charged Off – Loans where borrowers defaulted
    - Fully Paid – Loans where borrowers fully paid all dues
    - Current – Loans which are currently running
  - Total Loan Applications (Data available) : 39717
    - Charged Off + Fully Paid Loans : 38577
    - Currently running loans : 1140

Important Note :
In this presentation, details for Predictive Analytics have also been given (from Slide 17). We have built a predictive model which has been explained in the slides and the model aims at reducing the loss incurred by bank in deciding probable defaulting customers. Evaluation committee is requested to check the predictive analysis as well.

Data Analysis Methodology



DATA SOURCING

Data Understanding

Business Driven Derived Metrics

Data Driven Derived Metrics

Univariate Analysis

Segmented Univariate Analysis

Bivariate Analysis

Predict/Reduce Loss & Probable Loan Defaulters

Data Sourcing and Cleaning

- Data Understanding
  - CSV based single file data
  - Total Rows : 39717 records
  - Total Columns : 111 features
- Data Cleaning Methodology

Strategy	Remarks	Execution and Results	Number of columns found
Single Value Type Columns	Columns where count of unique values = 1	Remove columns which match strategy	60 (list attached as text file)
Columns with only 0s and NAs	Columns where count of unique value = 2 and it is only 0s and NAs	Remove columns which match strategy	3 "collections_12_mths_ex_med" "chargeoff_within_12_mths" "tax_liens"
Data Cleaning for different columns 1) int_rate 2) term 3) emp_length	1) int_rate – remove "%" and convert to numeric 2) term – remove "months" and convert to numeric 3) emp_length – remove "years/</>/+" & convert to numeric	Mutate columns with using Dplyr utility and update data frame	3 columns updated
Data Mutation – for date based columns 1) issue_d 2) earliest_cr_line 3) last_pymnt_d 4) next_pymnt_d 5) last_credit_pull_d	Convert all columns in Posix.Ct Date time object.	Mutate columns with using Dplyr and lubridate utility and update data frame	5 columns updated

Univariate & Segmented Univariate Analysis (1/6)

Inference :

G(01)

Annual Income does not show a consistent relationship with respect to loan status. ***Dropped from analysis***

G(02)

Home Ownership provides credible information about borrowers market value (Only 3 rows belong to “None” category). ***Should be considered for analysis***

G(03)

Verification status shows relationship with loan status, surprisingly verified loans seems to have more defaults. ***Should be considered for analysis***

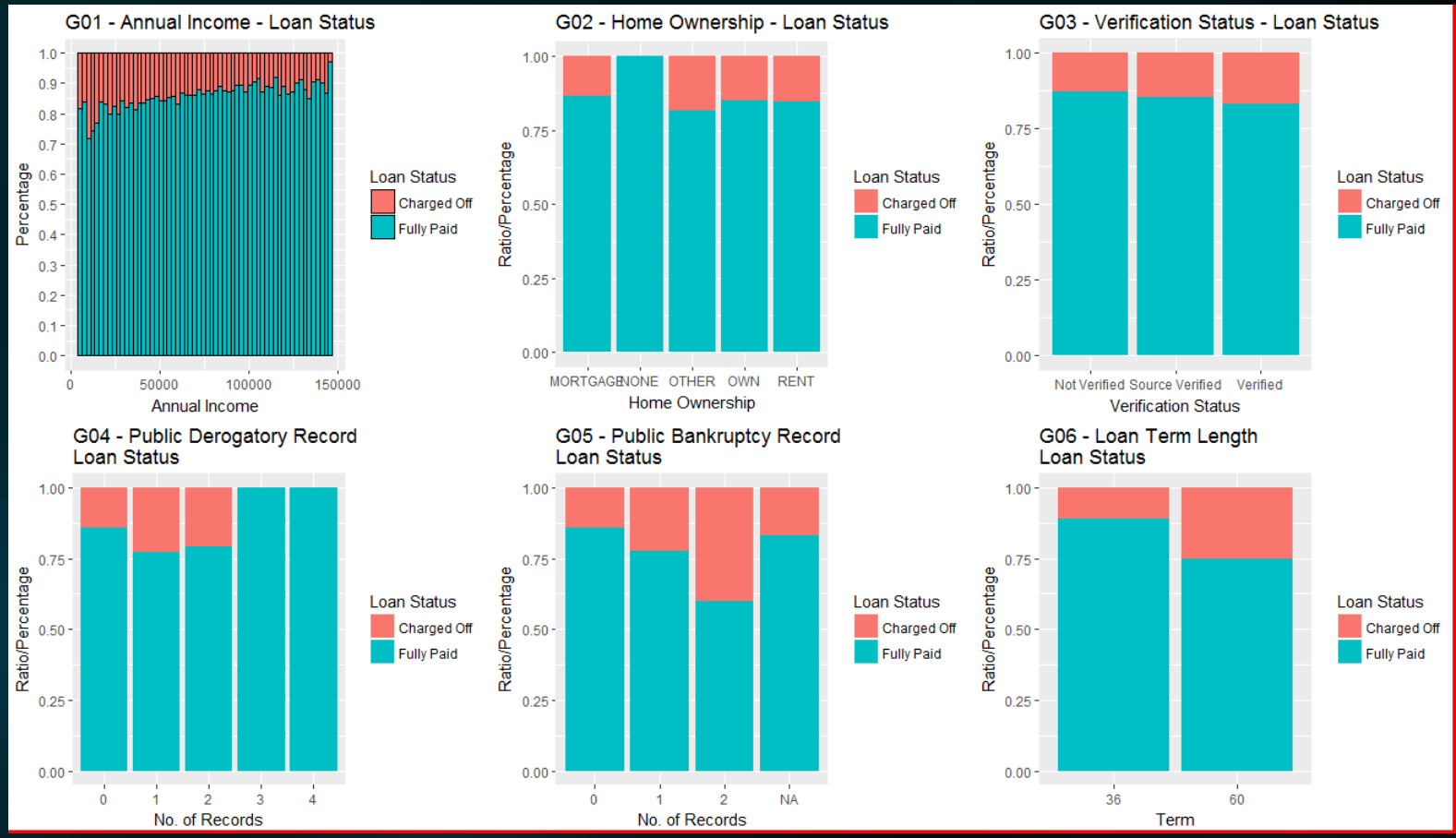
G(04) & G(05)

Public Derogatory Records does not show relationship with loan status, ***Dropped from analysis***

Logically as well as from data public bankruptcy records shows strong relationship with loan status . ***Should be considered for analysis***

G(06)

Loan term shows strong relationship with loan status. ***Should be considered for analysis***



Plot Methodology : Percentage distribution as per Loan Status

Aggregation : Various (as per plot)

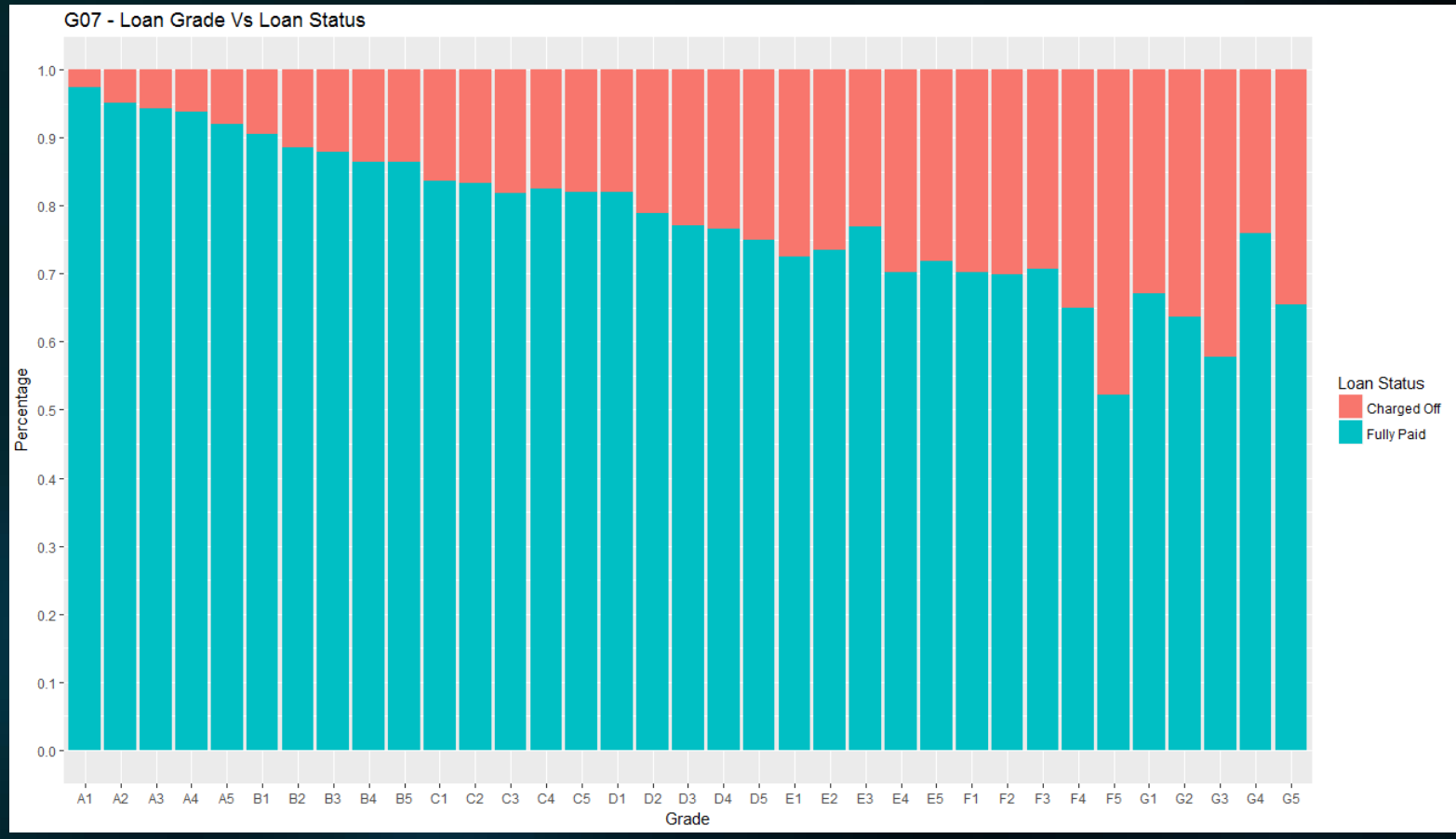
Univariate & Segmented Univariate Analysis (2/6) (Contd.)

Inference :

G(07)
Grade

Based on a loan application, a grade is assigned. This Grade/Subgrade is one of *most influential parameters* which not only effects the loan status but various other parameters also.

Grade/Subgrade of the loan shows a very strong relationship with loan status. Should be considered for analysis.



Plot Methodology : Percentage distribution as per Loan Status

Aggregation : Subgrade & Loan Status

Univariate & Segmented Univariate Analysis (3/6) (Contd.)

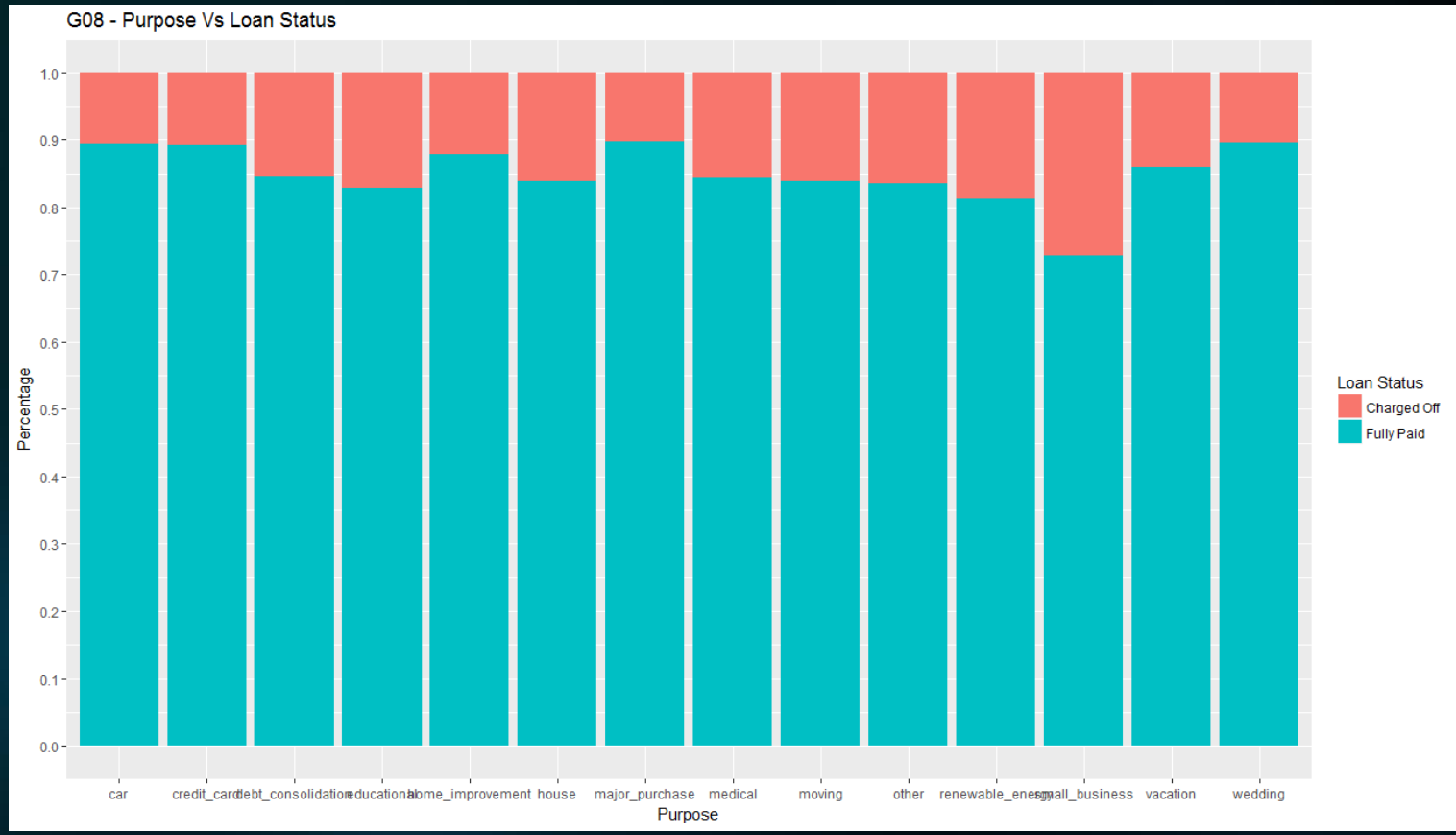
Inference :

G(08)

Purpose

Loan Purpose plays an important role and shows strong relationship with loan status. Maximum number of loans are taken for *debt_consolidation,* however, maximum number of defaults happen in “small businesses” ~ 25% of loans are defaulted

	purpose	cnt
1	car	1499
2	credit_card	5027
3	debt_consolidation	18055
4	educational	325
5	home_improvement	2875
6	house	367
7	major_purchase	2150
8	medical	681
9	moving	576
10	other	3865
11	renewable_energy	102
12	small_business	1754
13	vacation	375
14	wedding	926



Plot Methodology : Percentage distribution as per Loan Status

Aggregation : Purpose & Loan Status

Univariate & Segmented Univariate Analysis (4/6) (Contd.)

Inference :

G(09)

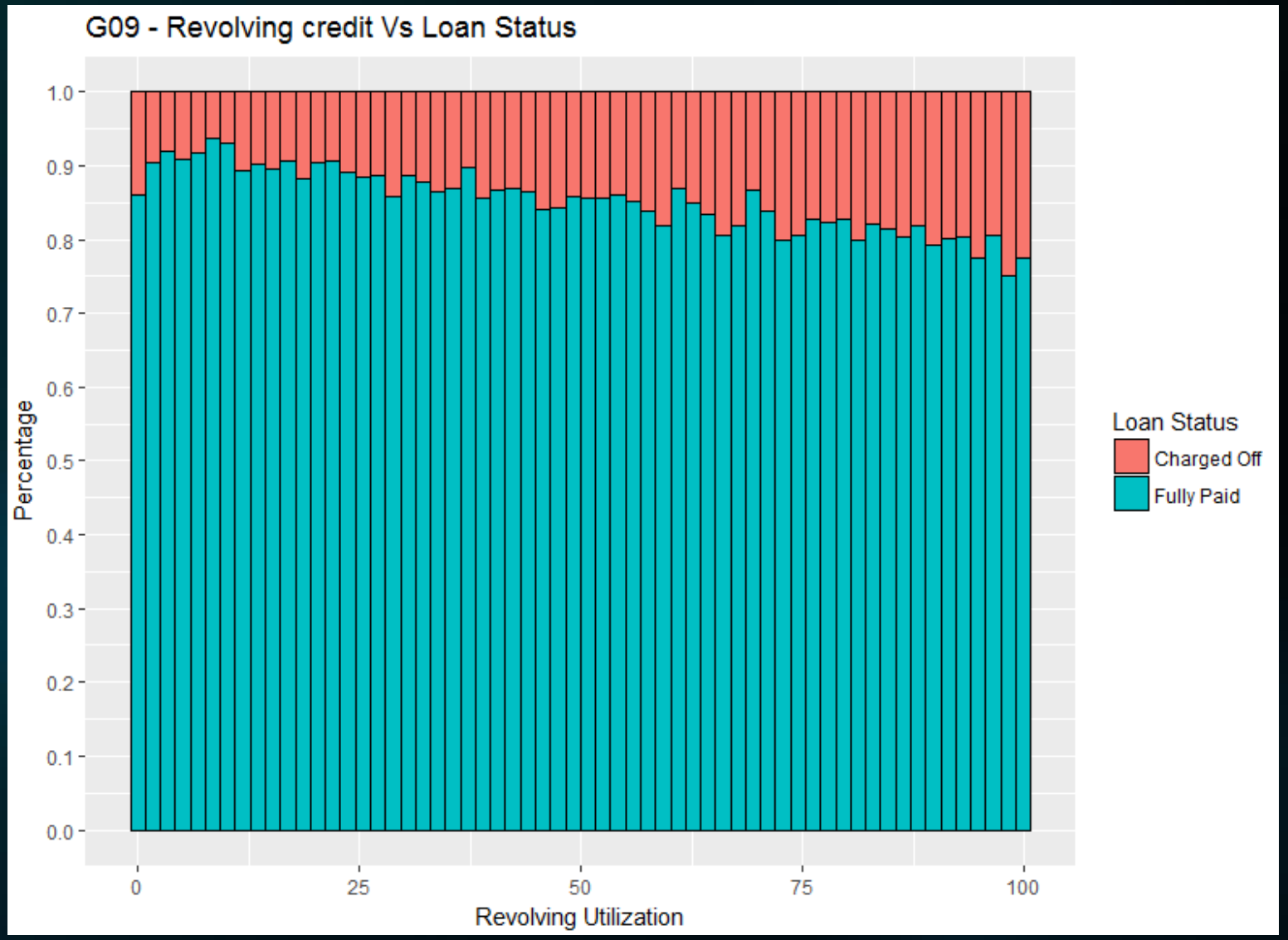
Revolving Credit :

"Revolving credit is a line of credit where the customer pays a commitment fee and is then allowed to use the funds when they are needed. It is usually used for operating purposes and can fluctuate each month depending on the customer's current cash flow needs. Revolving lines of credit can be taken out by corporations or individuals." – Source (Investopedia)

Revolving credit shows strong relationship with loan status. Should be considered for analysis.

Plot Methodology : Percentage distribution as per Loan Status

Aggregation : Revolving Credit & Loan Status



Univariate & Segmented Univariate Analysis (5/6) (Contd.)

Inference :

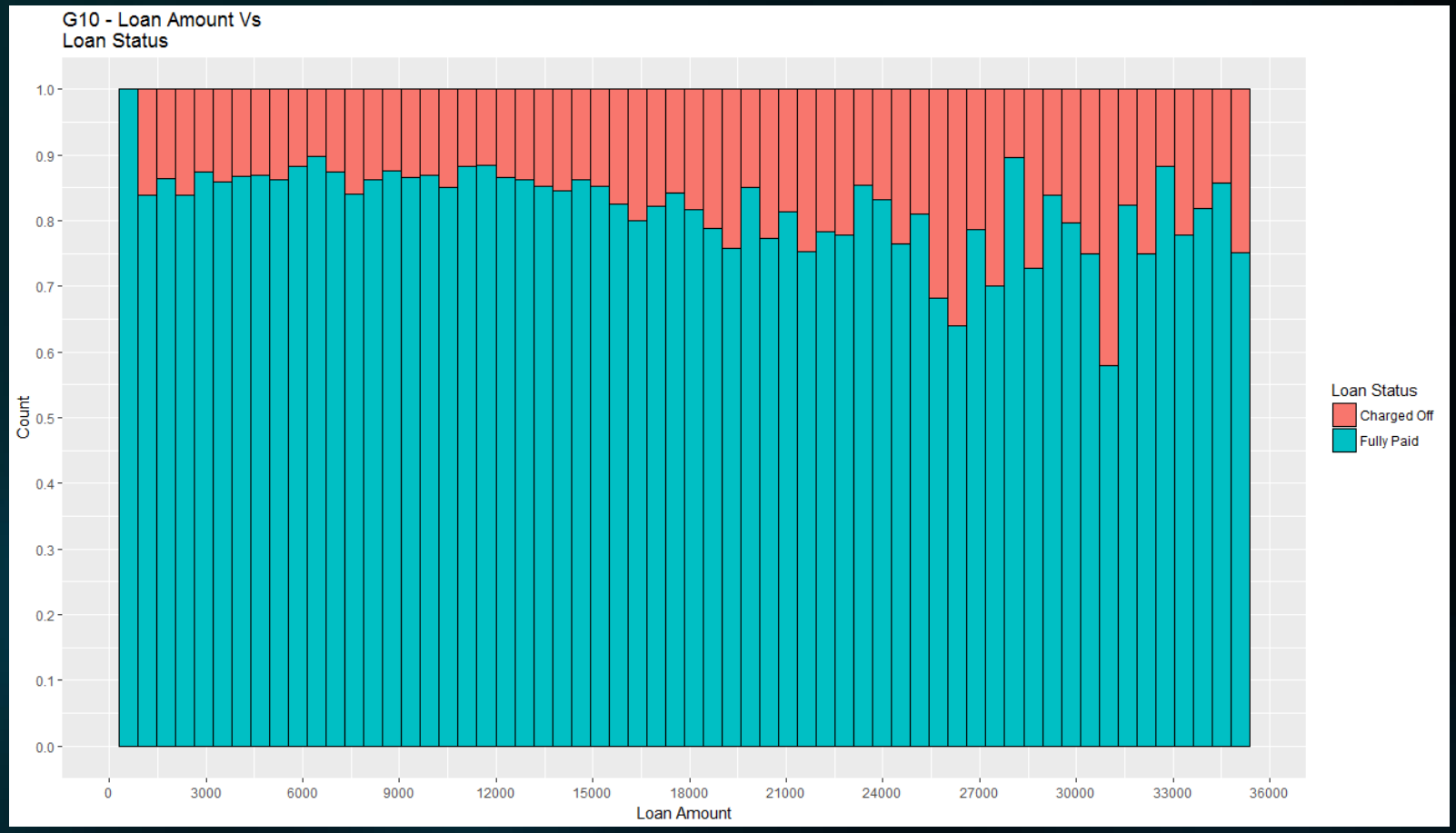
G(09)

Loan Amount

Loan amount shows a relationship with the loan status with the fact that higher loan amounts tend to be defaulted more than smaller loan amounts.

Due to the relationship inferred from the graph, loan_amount should be considered for analysis.

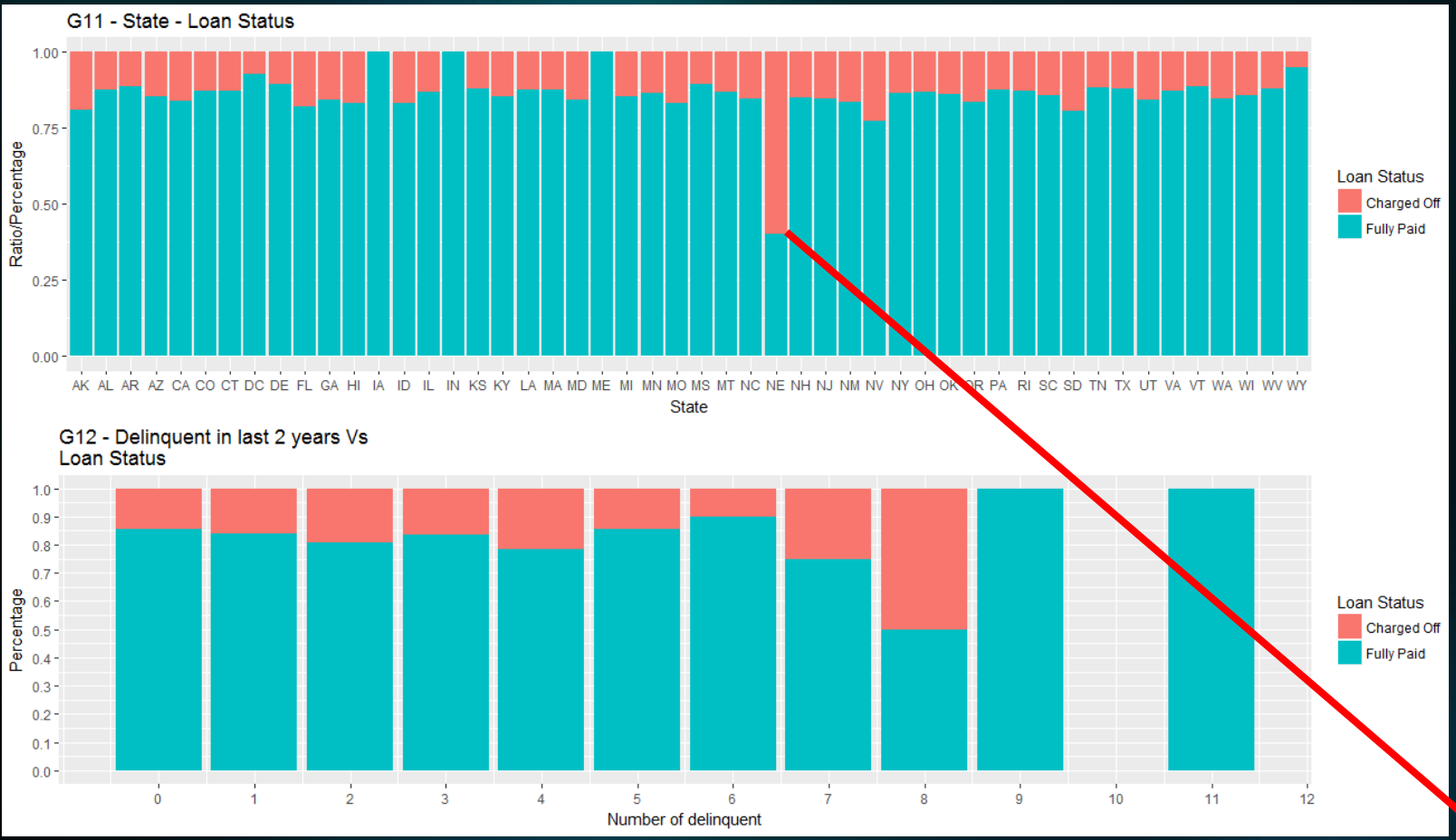
Note: The relationship of loan_amount and loan_status seems to be a bit inconsistent, therefore, we will be analyzing the loan_amount along with DTI ratio further for clearer understanding.



Plot Methodology : Percentage distribution as per Loan Status

Aggregation : Loan Amount & Loan Status

Univariate & Segmented Univariate Analysis (6/6) (Contd.)



Inference :

G(11)

Address State of the borrower shows an uneven relationship with loan status. Therefore, it is ***Dropped from analysis***

Maximum Loans from : California (CA) - 7099
Minimum Loans from : Nebraska (NE) - 5

G(12)

Logically delinquency should be an important factor, but data distribution and the pattern is uneven. Therefore, it is ***Dropped from analysis***

	delinq_2yrs	cnt
	<int>	<int>
1	0	35405
2	1	3303
3	2	687
4	3	220
5	4	62
6	5	22
7	6	10
8	7	4
9	8	2
10	9	1
11	11	1

Plot Methodology : Percentage distribution as per Loan Status

Aggregation : Various (as per plot)

NE (Nebraska) has very less number of loan accounts hence this outlier is of no value.

Bivariate Analysis - Correlation
Logical Loan Application Flow Process

Loan Application Flow Process (Logical)

When a borrower comes with a loan request, following methodology would have been followed:

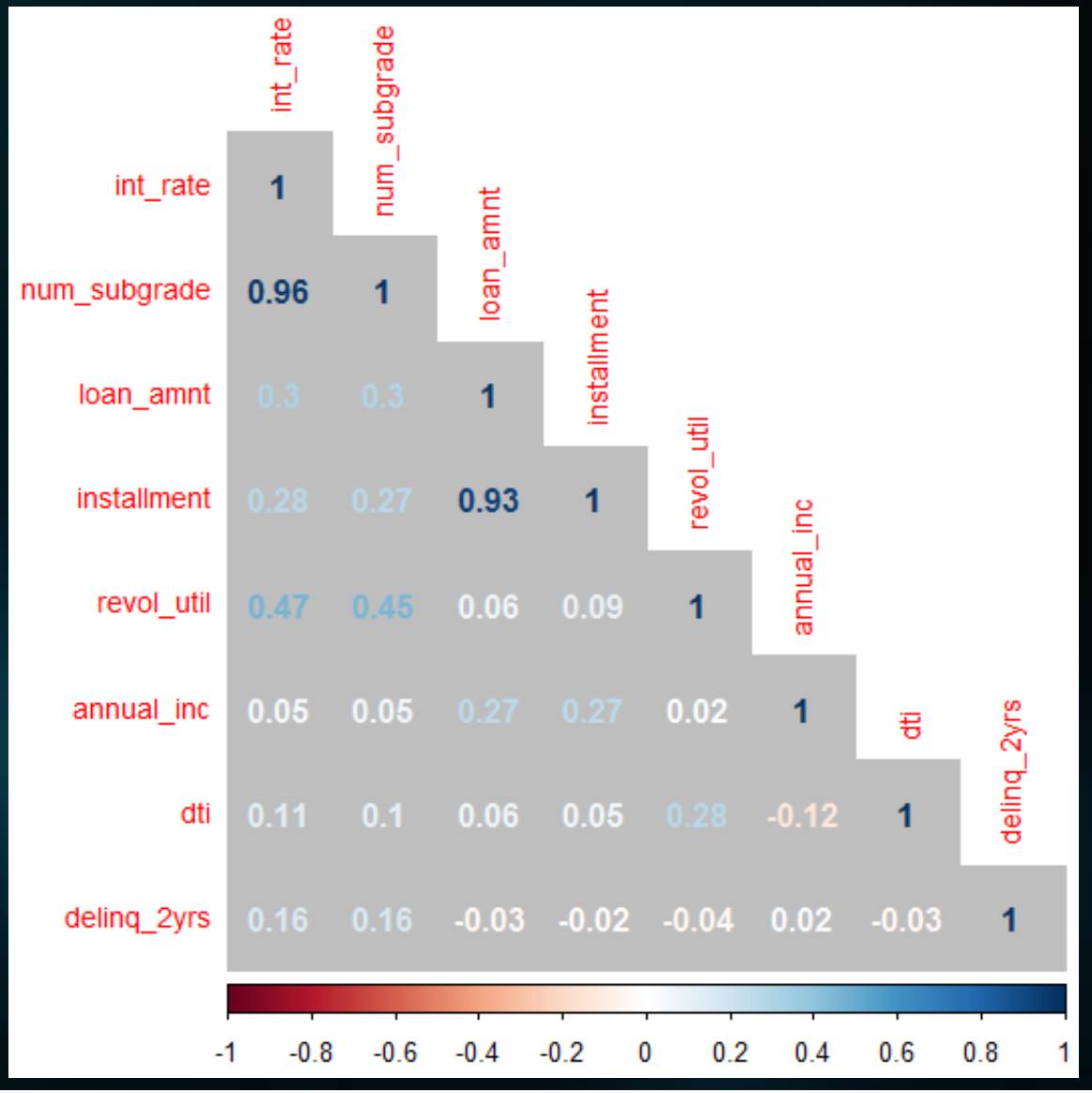
1) Based on the loan application submitted by the borrower, a specific grade is provided to the application
2) Based on the grade/sub_grade other parameters (such as interest rate etc.) are decided.

In order to decide influential parameters, it is important to check their relationship with grade/sub_grade as well.

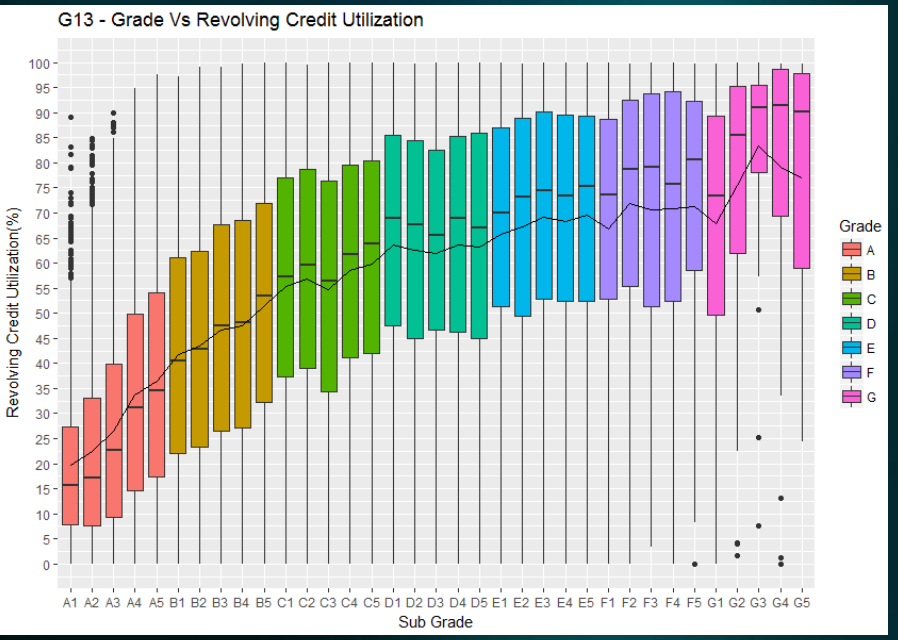
If a parameter's relation with grade/sub_grade is strong, then we may decide on considering only grade/sub_grade as the influencing parameter among the two parameters

Based on above thought process and co-relation matrix shown on the right, following conclusions can be drawn.

a) Sub_grade – Interest Rate have strong Co-relation
b) Loan_amnt – Installment have strong co-relation (Obvious)
c) Revol_util – int_rate & sub_grade seem to have a moderate co-relation

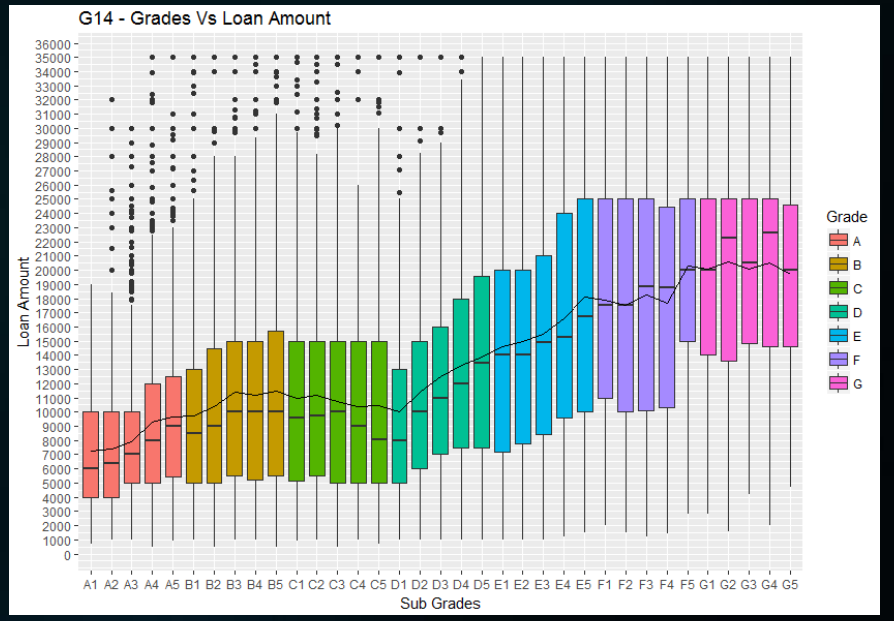


Bivariate Analysis



Plot Methodology : BoxPlot of Loan Amount
Aggregation : SubGrade and LoanStatus

Plot Methodology : BoxPlot of Revolving Util
Aggregation : SubGrade and LoanStatus



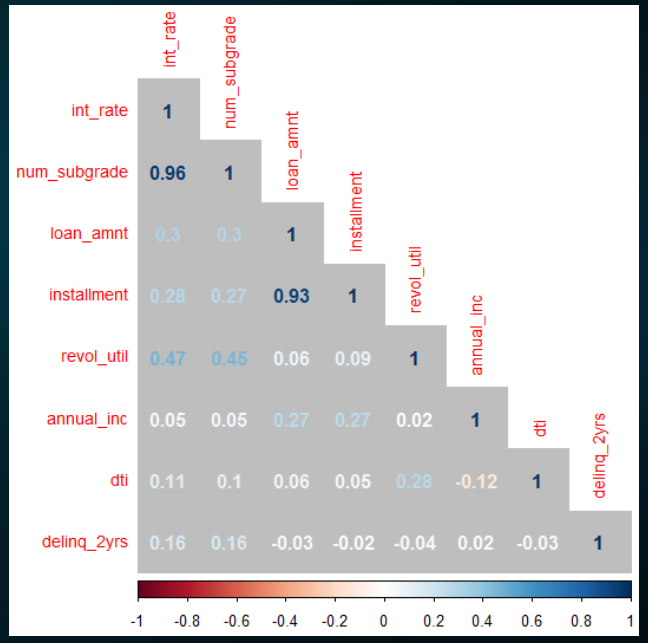
Inference :

G(13)

Grade/SubGrade Vs Revolving Credit Utilization

In line with the co-relation matrix, a moderate level of linkage can be seen between revol_util and grade/subgrade level. This co-relation seems to have exceptions, and hence binning should be done and revolving credit should be considered as an independent variable

Conclusion:
It would be better to consider revol_util as a separate variable and bin it for easier analysis. <- **DERIVED METRICS**



Inference :

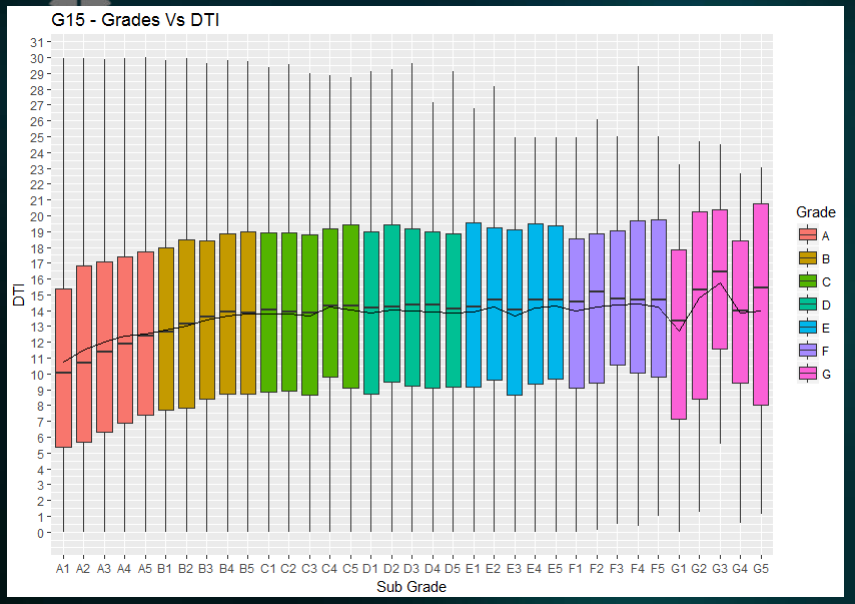
G(14)

Grade Vs Loan Amount

A certain level of co-relation can be seen between Loan Amount and Grade of the loan, however this co-relation does not seem to be consistent.

Conclusion:
It would better to consider Loan Amount as a separate variable, and bin the loan amount for easier analysis. <- **DERIVED METRICS**

Bivariate Analysis (Contd.)



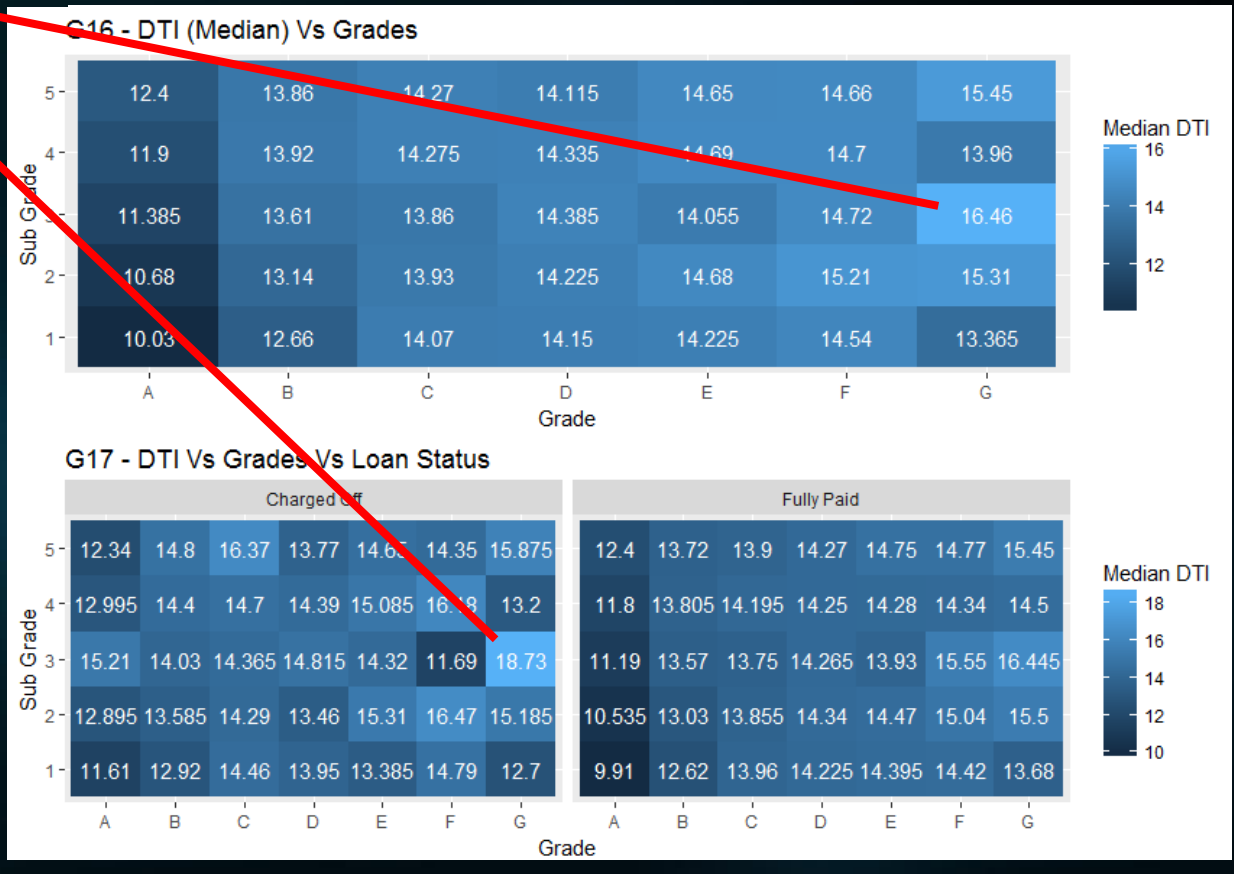
SubGrade "G3" Problem:

DTI Value (Median) for SubGrade G3 seems to be an outlier in the whole distribution. This aspect can be seen in Charged Off Loans as well as Full Paid Loans.

Loans given in G3 Grade should be looked in to carefully for probable "Charged Off Problems"

Plot Methodology : Tile Plot for DTI(Median)
Aggregation : Grade, SubGrade & LoanStatus

Plot Methodology : BoxPlot of DTI Vs SubGrade
Aggregation : SubGrade



Inference :

G(15,G16,G17)
Grade/SubGrade Vs DTI

DTI is the ratio calculated using the borrower's total monthly debt payments on the total debt obligations, divided by the borrower's self-reported monthly income.

From Tile Plot is clear that as the Sub Grade increases (A1 – G5) the DTI value is also increasing.

Conclusion:
DTI seems to be related to SubGrade hence we can take just Subgrade as a reference variable.

Bivariate Analysis (Contd.)
SubGrade “G3” Problem Description

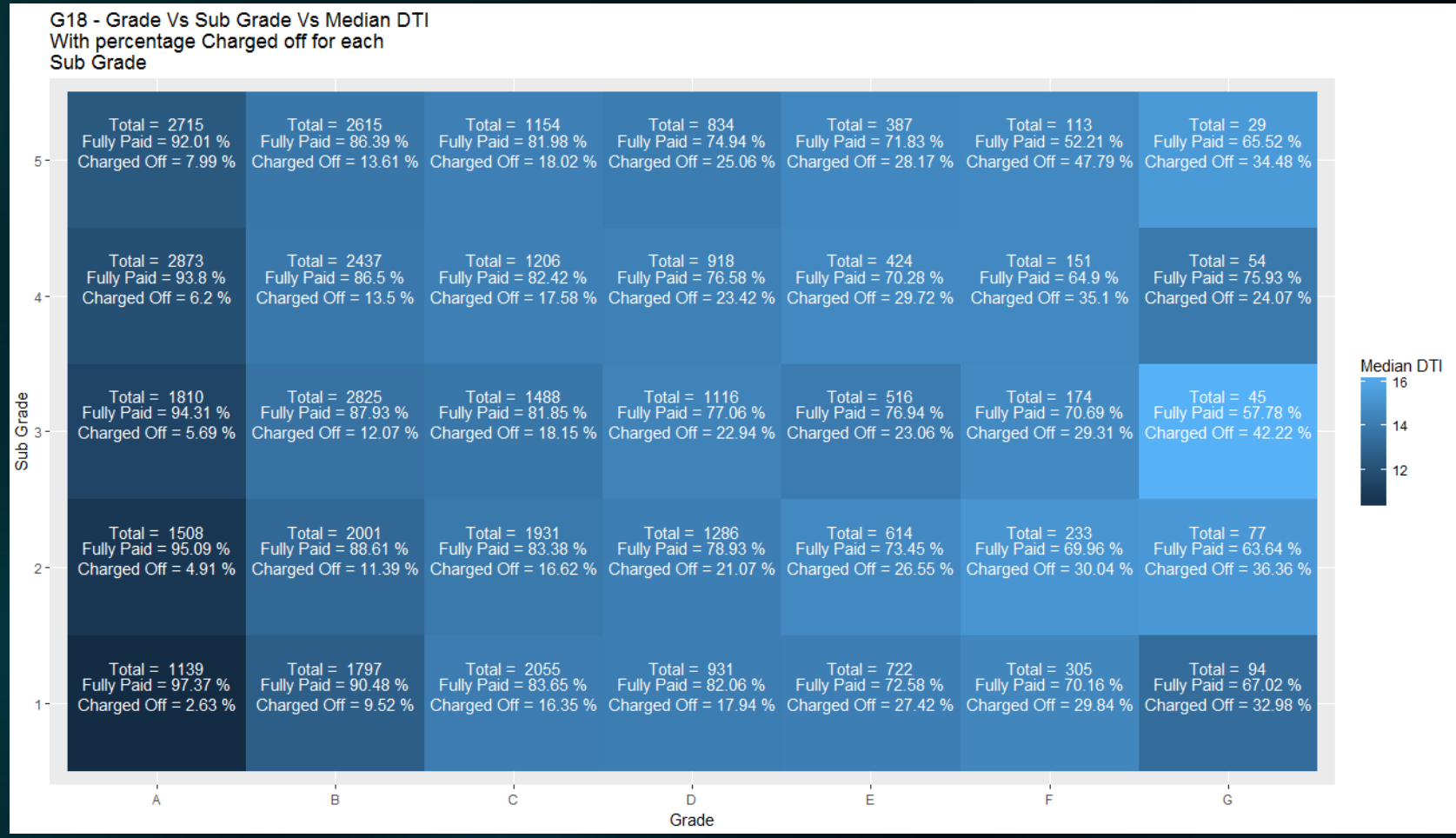
Inference :

G(18)

SubGrade “G3” Issue

There seems to be an issue with the loans being issued under SubGrade = “G3”

a) DTI value for loans issued under G3 seems to be exceptionally high, and does not follow an even pattern like other grades.
b) Percentage “Charged Off” loans under this category are high as well.
c) The banking institution should check the procedures being followed for giving away loans in SubGrade = “G3” category.

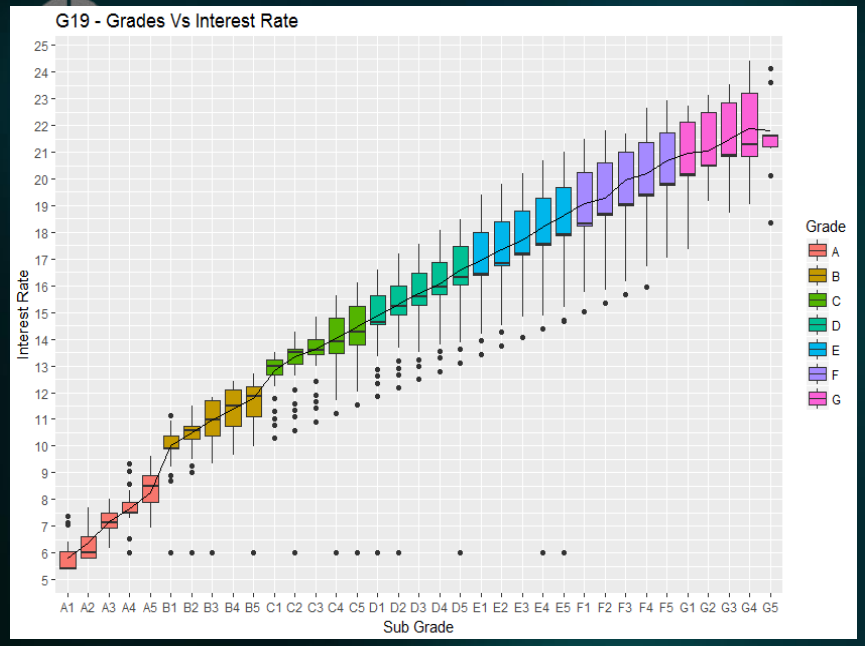


Plot Methodology : Tile Plot with median of DTI

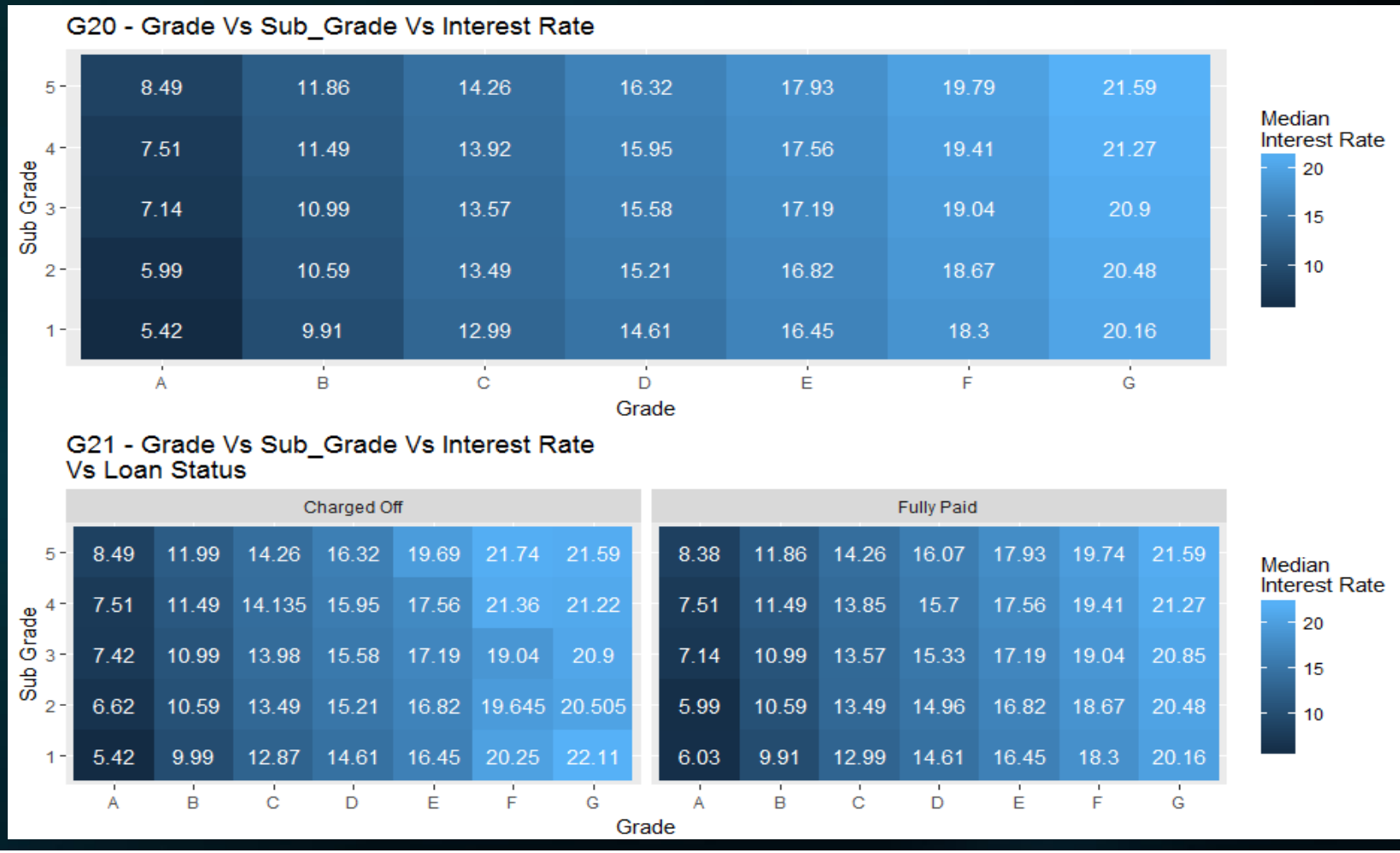
Aggregation : Grade, SubGrade Vs DTI(Median)

Bivariate Analysis (Contd.)

Plot Methodology : Boxplot Interest Rate against SubGrade
Aggregation : Grade, SubGrade Vs Interest Rate



Plot Methodology : Tile Plot with median of Interest Rate
Aggregation : Grade, SubGrade Vs Interest Rate



Inference :

G(19,20,21)

Grade/SubGrade Vs Interest Rate

Loan interest rate shows a clear co-relation with subgrade feature.

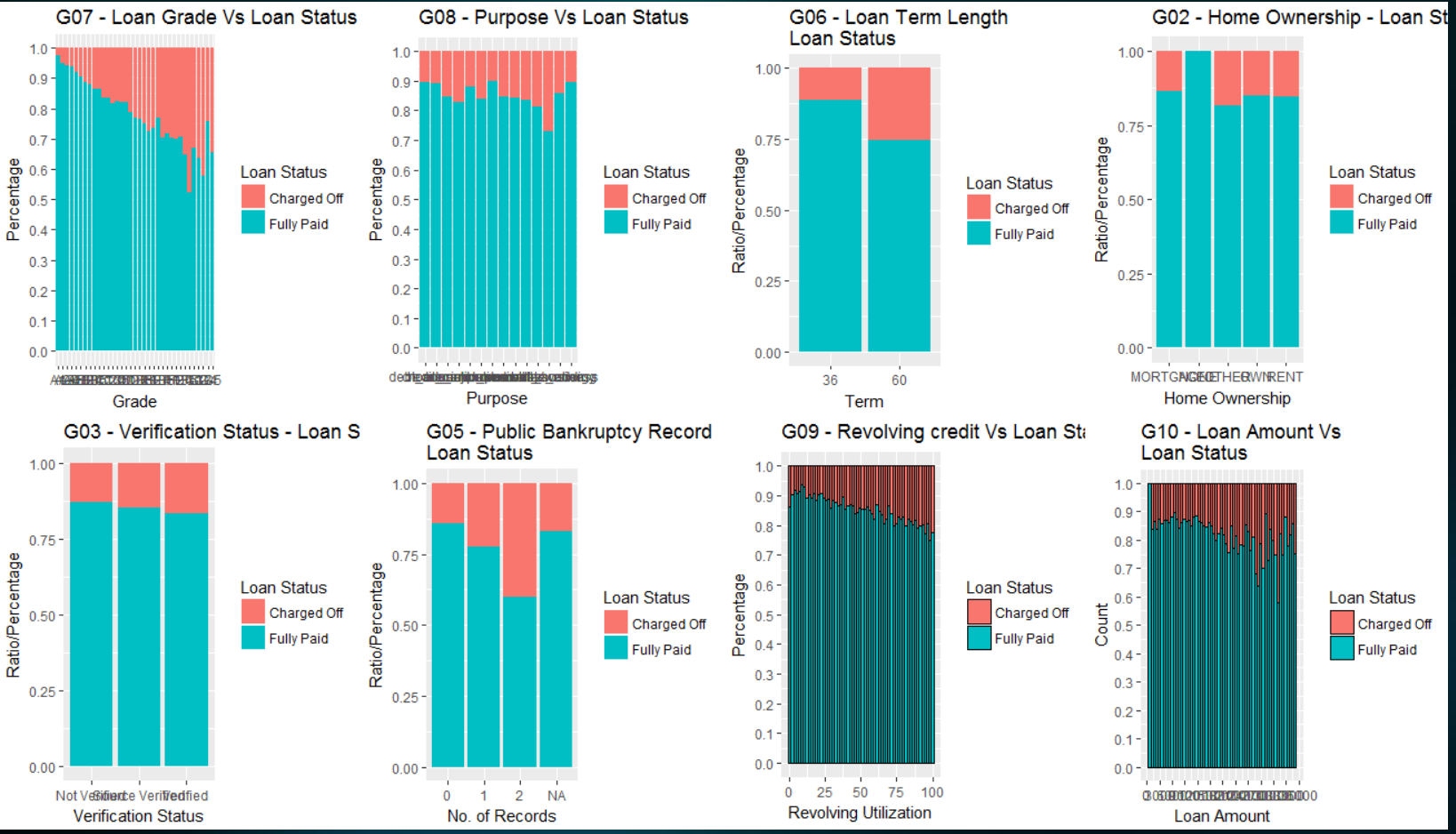
Conclusion:

Since we are considering grade as an influencing variable, therefore, interest rate would be considered invariably.

Hence we can leave away interest rate from being considered as an influencing variable.

Conclusion – EDA

Primary Influential Variables for Loan Status Analysis



Inference :

Based on EDA analysis, following variables can be said to be influencing the loan status to large extent.

Based on these parameters a predictive model can be created which will be able to predict the probability metrics as to whether currently running loans may end up as “Charged Off” or “Fully Paid”

Note: Below list has been created on the basis of level of influence.

1. sub_grade - G07
2. purpose - G08
3. term - G06
4. home_ownership - G02
5. verification_status - G03
6. public bankruptcy records - G05
7. revol_util_range - G09
8. loan_amnt_range - G10

Predictive Analysis

Logical Assumption taken for Predictive Analysis and Methodology

❑ Explanation :

Based on EDA shown above, we have zeroed down on 8 critical variables which effect the loan_status. The analysis below tries to predict the loans which are currently running, with a certain level of accuracy whether a loan might end up as "Charged Off" or "Fully Paid"

❑ Description of Analysis

Let us say there is a borrower who has following credentials

subGrade = G3
home_ownership = OWN
Verification status = Not Verified …. So and so forth.

❑ Analysis methodology

Step 1) We take up every such row which needs to be checked
Step 2) We take up the complete known data set where where loan_status is known
Step 3) We start by comparing looking for row/rows which match all the 8 conditions mentioned above
Step 4) There are possibilities that we might get to only 3 out of 8 conditions only, but that is not a problem
Step 5) Once we get to the closest matching conditions (whether 3,4,5...), we calculate the ratios
Step 6) We take the ratio = #Number of Charged Off Loans/#Number of total loans given * (level/9)
Step 7) Step (6) is the probability of borrower defaulting the loan

Probability that the member_id will result in a “Charged Off” loan or “Fully Paid” loan. Higher the ratio, higher is probability to default.

❑ Based on above process we are able to get following type of results.

id	member_id	sub_grade	purpose	term	home_ownership	verification_status	revol_util	loan_amnt	level	ratio	loan_status	loan_status_proj
1067874	1302235	B5	major_purchase	60	RENT	Source Verified	50.0	6000	7	0.87500000	Current	Charged Off
1034693	1264291	D4	debt_consolidation	60	RENT	Not Verified	87.1	16000	7	0.30078125	Current	Fully Paid
1046969	1277832	C1	debt_consolidation	60	MORTGAGE	Verified	24.1	11000	8	0.33333333	Current	Fully Paid

Cross Validation Of Predictive Analysis & Reduction of Loss To Bank

- ***Explanation :***

Based on the explanation provided previously, we have to decide on a “CutOff” probability so that

- Before the cutoff probability, all borrowers will be classified as “Safe/Good Customers” and their loans are predicted to be closed as “Fully Paid”
- Above the cutoff Probability, borrowers will be classified as “UnSafe/Bad Customers” and their loans are predicted to be closed as “Charged Off”

- ***How to decide the CutOff Probability***
  - We execute the same algorithm discussed above on existing data set where “Loan_Status” is known.
  - By performing above operation on multiple random sets of known data, we will get following results
    - Prediction Accuracy
    - For loans predicted as “Charged Off” who were actually “Fully Paid”, the loss is
      - Projected Business Loss of Good Customer
    - For Loans predicted as “Fully Paid” who were actually “Charged Off”, the loss is
      - Projected Business Loss to Defaulters
- We have to reach a “BALANCING VALUE”, which minimizes both the losses. The above analysis generates a data like this

cutOff	true_positive	true_negative	false_positive	false_negative	precision	recall	fscore	correct	incorrect	totalProspectiveBusinessLoss	totalDefaulterBusinessLoss
0.5	48	705	293	94	0.14076246	0.33802817	0.19875776	0.6605263	0.3394737	772352.1	363935.8
0.9	8	908	90	134	0.08163265	0.05633803	0.06666667	0.8035088	0.1964912	247456.9	509718.9
1	8	908	90	134	0.08163265	0.05633803	0.06666667	0.8035088	0.1964912	247456.9	509718.9

Prediction Accuracy

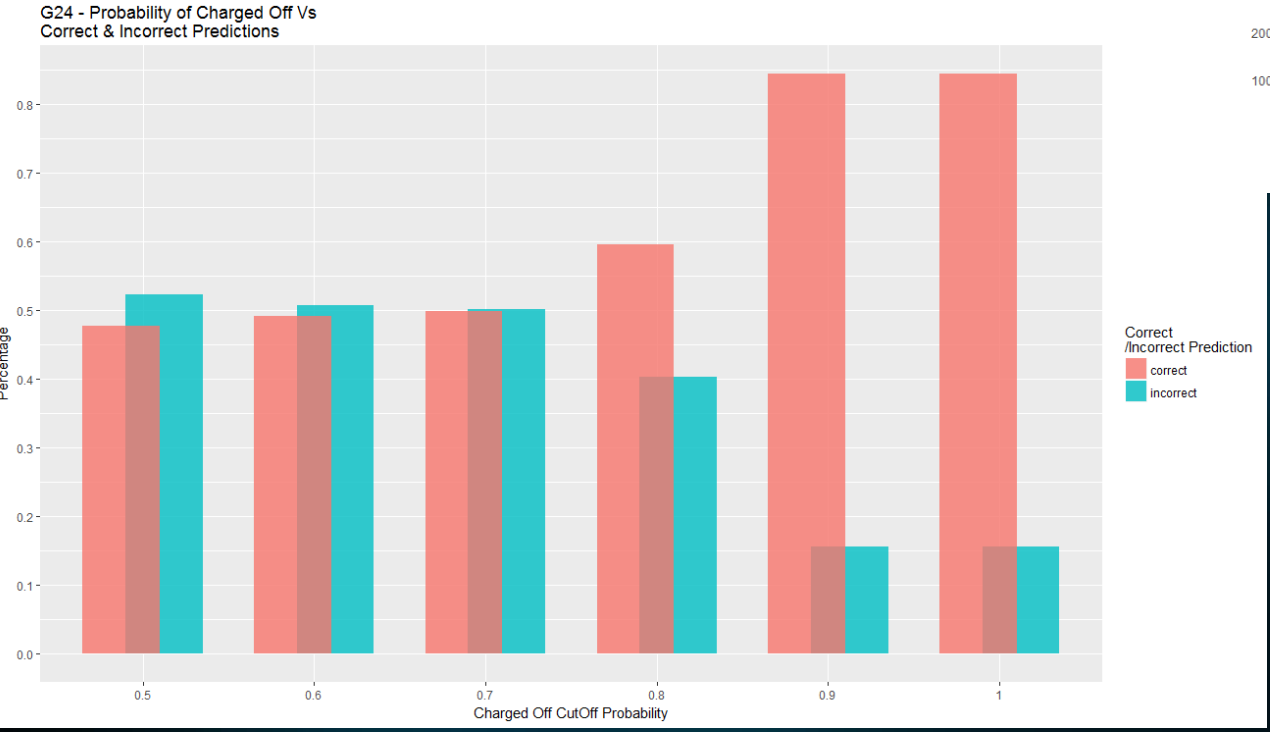
Projected Business Loss (in different cases)

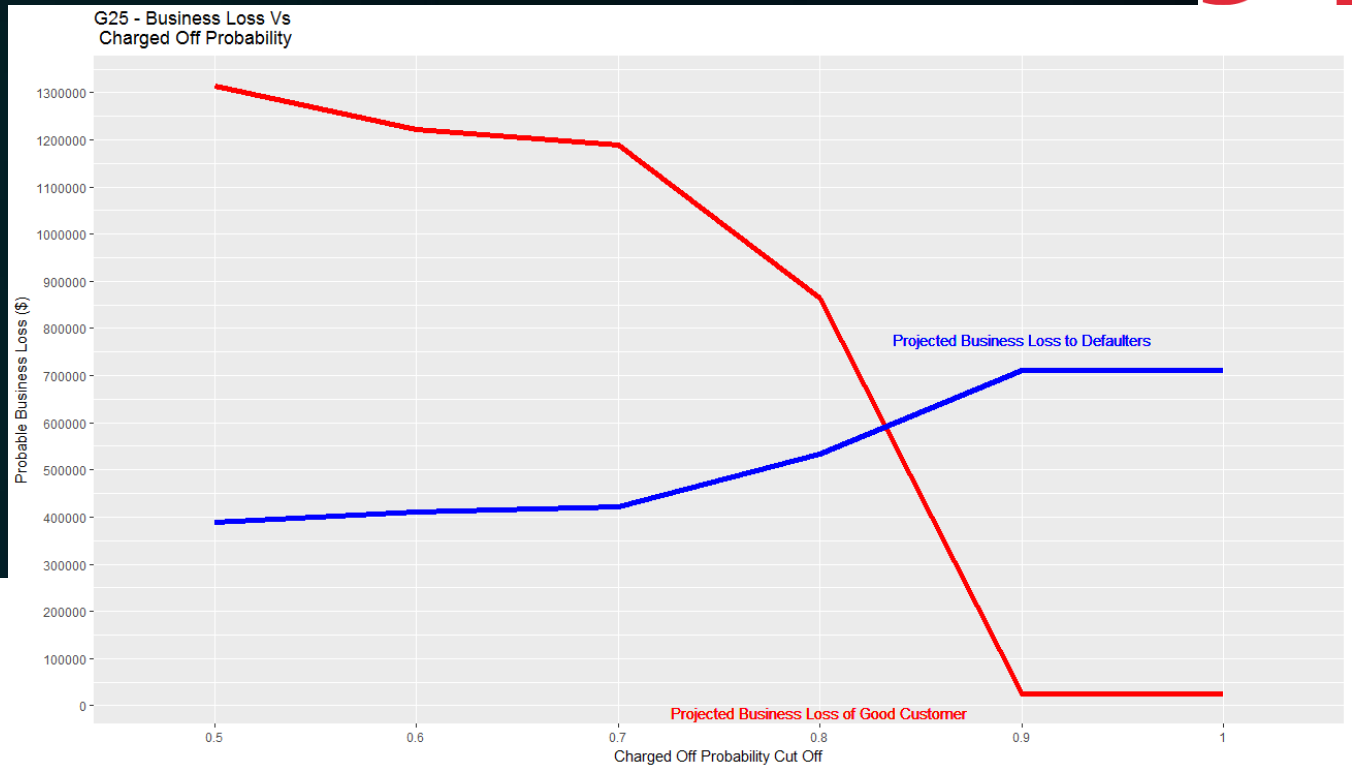
Predictive Analytics Results (Prediction Accuracy)

Inference :

For maximizing prediction accuracy, following can be chosen

- CutOff Probability for Charged Off categorization = 1
- But choosing the cutoff = 1 does not reduce the loss.





Inference :

Above graph shows the movement of losses incurred by the bank by choosing a specific “cutoff” value.

- It can be seen that lower cutoff value, increases the Good Customer Business loss but reduces loss to defaulters
- Similarly, higher cutoff value, increases the probable loss to defaulters, but reduces the Good Customer business loss.
- Hence, a balanced cutoff Value chosen ~ ***0.825***

Prediction for Currently Running Loans

- Based on the Predictive Analysis above
  - Out of total Loans with loan_status == Current
    - Total Loans = 1140 Loans
    - 302 Loans – *Charged Off (Projected)*
    - 838 Loans – *Fully Paid (Projected)*
- Above Prediction has been made through an CutOff probability value = ***0.825***
- The above prediction has been done to reduce the probable loss incurred by the bank.